Journal of Electronic Imaging

JElectronicImaging.org

Transformation-aware perceptual image metric

Petr Kellnhofer
Tobias Ritschel
Karol Myszkowski
Hans-Peter Seidel

Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, Hans-Peter Seidel, "Transformation-aware perceptual image metric," *J. Electron. Imaging* **25**(5), 053014 (2016), doi: 10.1117/1.JEI.25.5.053014.

Transformation-aware perceptual image metric

Petr Kellnhofer,[a,*] Tobias Ritschel,[a,b,c] Karol Myszkowski,[a] and Hans-Peter Seidel[a]
[a]Max-Planck-Institut für Informatik, Campus E1.4, Saarbrücken 66123, Germany
[b]Saarland University, Uni-Campus Nord, Saarbrücken 66123, Germany
[c]University College London, 66-72 Gower Street, London WC1E 6EA, United Kingdom

Abstract. Predicting human visual perception has several applications such as compression, rendering, editing, and retargeting. Current approaches, however, ignore the fact that the human visual system compensates for geometric transformations, e.g., we see that an image and a rotated copy are identical. Instead, they will report a large, false-positive difference. At the same time, if the transformations become too strong or too spatially incoherent, comparing two images gets increasingly difficult. Between these two extrema, we propose a system to quantify the effect of transformations, not only on the perception of image differences but also on saliency and motion parallax. To this end, we first fit local homographies to a given optical flow field, and then convert this field into a field of elementary transformations, such as translation, rotation, scaling, and perspective. We conduct a perceptual experiment quantifying the increase of difficulty when compensating for elementary transformations. Transformation entropy is proposed as a measure of complexity in a flow field. This representation is then used for applications, such as comparison of nonaligned images, where transformations cause threshold elevation, detection of salient transformations, and a model of perceived motion parallax. Applications of our approach are a perceptual level-of-detail for real-time rendering and viewpoint selection based on perceived motion parallax. [DOI: 10.1117/1.JEI.25.5.053014]

Keywords: image metric; motion; optical flow; homography; saliency; motion parallax.

Paper 16145P received Feb. 27, 2016; accepted for publication Aug. 30, 2016; published online Sep. 21, 2016.

1 Introduction

Models of human visual perception are an important component of image compression, rendering, retargeting, and editing. A typical application is prediction of differences in image pairs or detection of salient regions. Such predictions are based on the perception of luminance patterns alone and ignore that a difference might also be well explained by a transformation. As an example, the Hamming distance of the binary strings 1010 and 0101 is the same as between 1111 and 0000; however, the first pair is more similar in the sense of an edit distance, as 1010 is just a rotated, i.e., transformed version of 0101. We apply this idea to images, e.g., comparing an image and its rotated copy.

In current models of visual perception, transformation is not represented, leading to several difficulties. For image similarity or quality evaluation approaches, it is typically assumed the image pair is perfectly aligned (registered), which is directly granted in image compression, restoration, denoising, broadcasting, and rendering. However, in many other applications, such as visual equivalence judgement,[1] comparison of rendered and photographed scenes,[2] rephotography,[3] or image retargeting,[4] the similarity of images should be judged in the presence of distortions caused by transformations. Ecologically valid transformation[5] is a nonstructural distortion[6] and as such should be separated from others. However, current image difference metrics will report images that differ by such a transformation to be very dissimilar.[6] In the same vein, computational models of image saliency are based on luminance alone, or in the case of video, on the principle that motion has a "pop-up" effect.[7] However, for an image pair that differs by a spatially varying transformation some transformations might be more salient, not because they are stronger, but because they are distinct from others. Finally, motion parallax is compensated for easily and not perceived as a distortion but as a depth cue (Ref. 8, Ch. 28). We will show that all the difficulties in predicting the perception of transformed images can be overcome by an explicit model of human perception of transformations such as we propose.

In this work, we assume the optical flow[5] of an image pair to be given, either by producing it using three-dimensional (3-D) graphics or (typically with a lower precision) using computer vision techniques and focus on how the human visual system (HVS) represents transformations. We decompose the flow field into a field of elementary transformations,[9] a process that is likely to also happen in the dorsal visual pathway of the primate brain.[10] From this representation, we can model the effect of transformations on the perception of images. For comparing images, strong or incoherent transformations generally make the perception of differences increasingly difficult. We model this effect using a measure of transformation entropy. When given an image pair that differs by a transformation, we predict where humans will perceive differences and where not (Fig. 1). Using our representation, we can compare transformations and predict which transformations are salient compared to others. Finally, spatially varying transformations result in motion parallax, which can serve as a depth cue.

*Address all correspondence to: Petr Kellnhofer, E-mail: pkellnho@mpi-inf.mpg.de

In this work, we make the following contributions:

- A perceptually motivated decomposition of optical flow.
- A transformation-aware image difference metric.
- Prediction of transformation saliency.
- Estimation of motion parallax as a depth cue.

2 Background

In this section, we review the perceptual background of our approach. We will recall the idea of mental transformation and its relation to optical flow, saliency, as well as the basics of entropy in human psychology. The discussion of previous work for the two main applications we propose (image differences, saliency), is found in Sec. 4.

2.1 *Mental Transformation*

Mental transformations of space play an important role in everyday tasks such as object recognition, spatial orientation, and motion planning. Such transformations involve both objects in the space as well as the egocentric observer position. Mental rotation is the best understood mental transformation,[11] where the time required to judge the similarity between a pair of differently oriented objects of arbitrary shape is proportional to the rotation angle both for two-dimensional (2-D) (image plane) and 3-D rotations, irrespective of the chosen rotation axis. Similar observations have been made for the size scaling and translation (in depth),[12] where the reaction time is shorter than for rotation. Moreover, in combined scaling and rotation[13] as well as translation and rotation[12] the response time is additive with respect to each component transformation. This may suggest that there are independent routines for such elementary transformations, which jointly form a procedure for handling any sort of movement that preserves the rigid structure of an object.[12] Another observation is that the mental transformation passes through a continuous trajectory of intermediate object positions, not just the beginning and end positions.[14]

A more advanced mental transformation is perspective transformation.[15] From our own experience, we know that observing a cinema screen from a moderately off-angle perspective does not reduce perceived quality, even if the retinal image underwent a strong transformation. One explanation for this ability is that humans compensate for the perspective transformation by mentally inverting it.[16]

Apparent motion in the forward and backward directions is induced when two 3-D-transformed (e.g., rotated) copies of the same object are presented alternatively at proper rates. As the transformational distance (e.g., rotation angle) increases, the alternation rate must be reduced to maintain the motion illusion. Again, this observation strongly suggests that the underlying transformations require time to go through intermediate stages, a 3-D representation is utilized internally,[17] and elementary transformations are individually sequential-additive.[18]

The HVS is able to recover depth and rigid 3-D structure from two views (e.g., binocular vision and apparent motion) irrespective whether the perspective or orthographic projection is used, and adding more views has little impact.[19] This indicates that the HVS might use some perceptual heuristics to derive such information as the structure-from-motion theorem stipulates that at least three views are needed in the case of orthographic projection (or under weak perspective).[20]

The 3-D internal representation in the HVS and the rigidity hypothesis in correspondence finding, while tracking moving objects, is still a matter of scientific debate. Eagle et al.[21] have found a preference toward translation in explaining competing motion transformations in a two-frame sequence with little regard for the projective shape transformations.

2.2 *Optical Flow*

The idea of optical flow dates back to Gibson[5] and has become an essential part of computer vision and graphics where it is mostly formalized as a 2-D vector field that maps locations in one image to locations in a second image, taken from a different point in time or space. Beyond the mapping from points to points, Koenderink[9] conducted a theoretical analysis of elementary transformations, such as expansion/contraction (radial motion), rotation (circular motion), and sheer (two-component deformation), which can be combined with translation into a general affine transformation. Such transformations map differential area to differential area. Electrophysiological recordings have shown that specialized cells in the primate brain are selective for each elementary transformation component alone or combined with translation[10] (refer also to Ref. 8, Ch. 5.8.4). A spatially varying optical flow field does not imply a spatially varying field of transformations: a global rotation that has small displacements in the center and larger displacements in the periphery can serve as an example. For this reason, our perceptual model operates on a field of elementary transformations computed from homographies instead of a dense optical flow. Homography estimation is commonly used in the video-based scene 3-D analysis, and the best results are obtained when multiple views are considered.[20]

In computer graphics, the use of elementary transformation fields is rare, with the exception of video stabilization and shape modeling. In video stabilization, spatially varying warps of handheld video frames into their stabilized version are performed with a desired camera path. Typically a globally reconstructed homography is applied to the input frame, before the optimization-driven local warping is performed,[22] which is conceptually similar to our local homography decomposition step (Sec. 3.2). Notably, the concept of subspace stabilization[23] constructs a lower-dimensional subspace of the 2-D flow field, i.e., a space with a lower number of different flows, i.e., lower entropy. In shape modeling, flow fields are decomposed into elementary transformations to remove all but the desired transformations, i.e., to remain as-rigid-as-possible when seeking to preserve only rotation.[24]

2.3 *Visual Attention*

Moving objects and "pop-out" effects are strong attractors of visual attention.[7] The classic visual attention model proposed by Itti et al.[25] apart from the common neuronal features, such as intensity contrast, color contrast, and pattern orientation can handle also four oriented motion energies (up, down, left, and right). Differently, in our work, we detect saliency of motion, which pops out not just because it is present and the rest is static, but because it is different from other motion in the scene, such as many rotating objects where one rotates differently. As humans understand motion in form

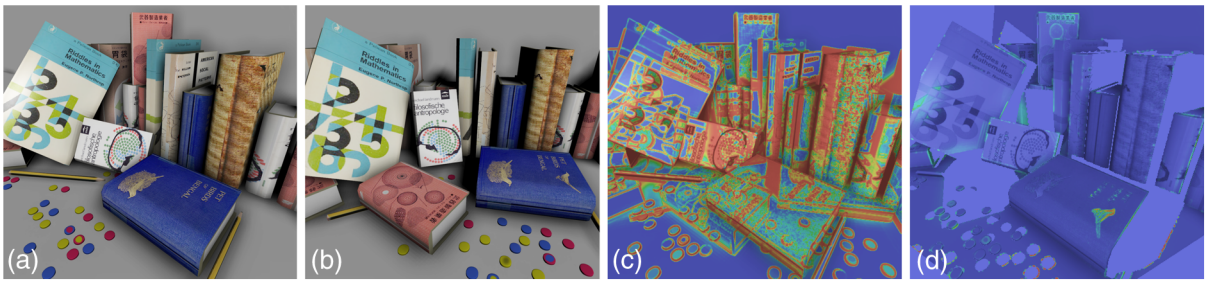

Fig. 1 Given input image (a) that underwent deformations producing image (b), common perceptual image metrics report unusable results (c) as they do not account for the HVS's ability to compensate for transformations. Our transformation-aware approach models the ability to compensate for transformations (d) and its limitations when transformations are too strong (red book) or too incoherent (chips).

of elementary transformations,[10] our analysis is needed to find those differences.

2.4 *Motion Parallax*

For a moving viewer, objects at one depth undergo a different flow than objects in their spatial neighborhood at different depth. This effect, called "motion parallax," is both a depth cue[26] and a grouping Gestalt cue.[27] In relation with translation and the four elementary transformations over the flow field as derived by Koenderink,[9] the corresponding motion parallax components can be distinguished: linear motion, expansion or contraction, rotation, shear-deformation, and compression- or expansion-deformation parallax (Ref. 8, Ch. 28.1.3). We propose a motion parallax measure, which approximates each of those components, although in our applications we found that the linear motion parallax plays the key role.

2.5 *Entropy*

Information entropy is a measure of complexity in the sense of how much a signal is expected or not.[28] If it is expected, the entropy is low, otherwise it is high. In our approach, we are interested in the entropy of transformations, which tells apart uniform transformations from incoherent ones, such as disassembling a puzzle. Assembling the puzzle is hard, not because the transformation is large, but because it is incoherent, i.e., it has a high entropy. This view is supported by studies of human task performance:[29,30] Sorting cards with a low entropy layout can be performed faster than sorting with high entropy. In computer graphics, entropy of luminance is used for the purpose of alignment,[31] best-view selection,[32] light source placement,[33] and feature detection but was not yet applied to transformations.

3 Our Approach

3.1 *Overview*

Our system consists of two layers (Fig. 2). A model layer described in this section and an application layer, described in Sec. 4. Input to the model layer are two images where the second image differs from the first one by a known mapping, which is assumed to be available as a spatially dense optical flow field. This requires either use of optical flow algorithms that support large displacements[34] and complex mappings[23] or use of computer-generated optical flow (a.k.a. motion field). Output of our method is a field of perceptually scaled elementary transformations and a field of transformation entropy ready to be used in different applications.

Our approach proceeds as follows (Fig. 2). In the first step (Sec. 3.2), we convert the optical flow field that maps positions to positions into an overcomplete field of local homographies, describing how a differential patch from one image is mapped to the other image. While classic flow only captures translation, the field of homographies also captures effects such as rotation, scaling, shear, and perspective. Next, we factor the local homographies into "elementary" translation, scaling, rotation, shear, and perspective transformations (Sec. 3.3). Also, we compute the local entropy of the transformation field, i.e., how difficult it is to understand the transformation (Sec. 3.4). Finally, the magnitude of elementary transformations is mapped to scalar perceptual units, such that the same value indicates roughly the same sensitivity (Sec. 3.5).

Using the information above allows for several applications. Most importantly, we propose an image difference metric (Sec. 4.1) that is transformation-aware. We model the threshold elevation, which determines how much the smallest perceivable difference between two images increases as

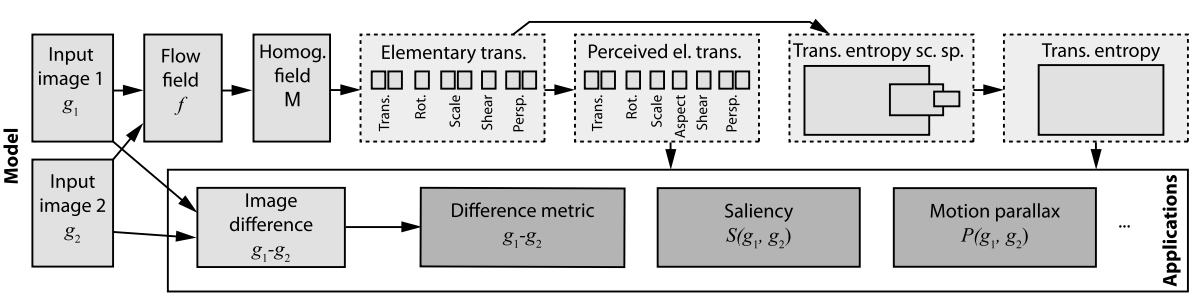

Fig. 2 Flow of our approach

a function of transformation strength and complexity, i.e., entropy. The second application is a visual attention model that can detect what transformations are salient (Sec. 4.2). Finally, the amount of perceived parallax in the image pair can be computed from the above information (Sec. 4.3).

3.2 *Homography Estimation*

Input is two images with luminances g_1 and $g_2(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}$ and $\mathbf{x}$ as spatial location as well as a flow $f(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}^2$ from g_1 to g_2. First, the flow field is converted into a field of homography transformations.[20] A homography maps a differential image patch into the second image while optical flow maps single pixel positions to other pixel positions. In human vision research, this Helmholtz decomposition was conceptually proposed by Koenderink[9] and later confirmed by physiological evidence.[10] Examples of homographies are shown in Fig. 3(a). In our case, homographies are 2-D projective 3×3 matrices. While 2×3 matrices can express translation, rotation, and scaling, the perspective component allows for perspective foreshortening.

We estimate a field of homographies, i.e., a map that describes for every pixel where its surrounding patch is going. We compute this field $\mathbf{M}(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}^{3\times3}$ by solving a motion discontinuity-aware moving least-squares problem for every pixel using a normalized eight-point algorithm.[35] The best transformation $\mathbf{M}(\mathbf{x})$ in the least squares sense minimizes

$$\int_{\mathbb{R}^2} w(\mathbf{x},\mathbf{y}) \left\| f(\mathbf{y}) - \phi\left[\mathbf{M}(\mathbf{x}) \begin{pmatrix} \mathbf{y} \\ 1 \end{pmatrix}\right] \right\|_2^2 \mathrm{d}\mathbf{y}, \tag{1}$$

where $\phi(\mathbf{v}) = (v_1/v_3, v_2/v_3)^T$ is a homogeneous projection and

$$w(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma_\mathrm{d}) \exp(-\|f(\mathbf{x}) - f(\mathbf{y})\|_2^2/\sigma_\mathrm{r}) \tag{2}$$

is a bilateral weight function[36] that accounts more for locations that are spatially close (domain weight) and have a similar flow (range weight). The parameters σ_r and σ_d control the locality of the weight. The range-weighting assures to not mix different flows into one wrong estimate of the homography, but to keep them separate [Fig. 3(b)] resulting in a pixel-accurate, edge-aware field.

In the discrete case of Eq. (1), for pixel $\mathbf{x}$ we find one $\mathbf{M}$ that minimizes

$$\sum_{i\in\mathcal{N}} w_i \left\| \mathbf{f}^i - \phi\left[\mathbf{M}\begin{pmatrix} \mathbf{y}^i \\ 1 \end{pmatrix}\right] \right\|_2^2, \tag{3}$$

where $\mathcal{N}$ is a 5×5 neighborhood around pixel location $\mathbf{x}$, and $\mathbf{f}^i$ and w_i are the flow and the bilateral weight of neighbor pixel i.

We solve this as a homogenenous linear least squares problem in form $\mathbf{Bm} = 0$. For one flow direction $\mathbf{f}_i$ at position $\mathbf{y}_i$ and a matrix $\mathbf{M}$, we require

$$\mathbf{f}_i - \phi\begin{pmatrix} y_1^i m_{11} + y_2^i m_{12} + m_{13} \\ y_1^i m_{21} + y_2^i m_{22} + m_{23} \\ y_1^i m_{31} + y_2^i m_{32} + m_{33} \end{pmatrix} = 0,$$

which after expanding $\phi(\mathbf{v})$ and rewriting vectors into two equations turns into

$$f_1^i - (y_1^i m_{11} + y_2^i m_{12} + m_{13})/(y_1^i m_{31} + y_2^i m_{32} + m_{33}) = 0$$
$$f_2^i - (y_1^i m_{21} + y_2^i m_{22} + m_{23})/(y_1^i m_{31} + y_2^i m_{32} + m_{33}) = 0,$$

which can be converted into a linear form by multiplying by the denominator

$$y_1^i f_1^i m_{31} + y_2^i f_1^i m_{32} + f_1^i m_{33} - y_1^i m_{11} - y_2^i m_{12} - m_{13} = 0$$
$$y_1^i f_2^i m_{31} + y_2^i f_2^i m_{32} + f_2^i m_{33} - y_1^i m_{21} - y_2^i m_{22} - m_{23} = 0.$$

We write $\mathbf{a}_1^{i\mathrm{T}}\mathbf{m} = 0$ and $\mathbf{a}_2^{i\mathrm{T}}\mathbf{m} = 0$ with

$$\mathbf{m} = (m_{11}, m_{12}, m_{13}, m_{21}, m_{22}, m_{23}, m_{31}, m_{32}, m_{33})^\mathrm{T}$$
$$\mathbf{a}_1^i = (-y_1^i, -y_2^i, -1, 0, 0, 0, f_1^i y_1^i, f_1^i y_2^i, f_1^i)^\mathrm{T}$$
$$\mathbf{a}_2^i = (0, 0, 0, -y_1^i, -y_2^i, -1, f_2^i y_1^i, f_2^i y_2^i, f_2^i)^\mathrm{T}.$$

This way, for every neighbor i, we compute a vector $\mathbf{a}_{\{1,2\}}^i$, $i \in (1, |\mathcal{N}|)$. Let $\mathbf{b}_{\{1,2\}}^i = \mathbf{w}_i \mathbf{a}_{\{1,2\}}^i$ be a weighted version of the error vector and $\mathbf{B}$ the 9×50 matrix that stacks all those error vectors $\mathbf{b}_{\{1,2\}}^i$

$$\mathbf{B} = (\mathbf{b}_1^1, \mathbf{b}_2^1, \mathbf{b}_1^2, \mathbf{b}_2^2, \mathbf{b}_1^3, \mathbf{b}_2^3, \cdots \mathbf{b}_1^{25}, \mathbf{b}_2^{25})^\mathrm{T}.$$

Finally, the homography $\mathbf{m}$ that minimizes $\|\mathbf{Bm}\|_2^2$ is found by solving a homogeneous linear system (HLS) of the form $\mathbf{Bm} = \mathbf{0}$. Pseudoinversion would only lead to trivial solution for HLS, therefore it cannot be used to solve the problem. Instead singular value decomposition in combination with preconditioning by translation of matched areas to the origin and normalization of their scale is commonly used.[35]

The procedure is similar to fitting of a single homography in computer vision.[20] It is more general, as our flow field is not explained by a rigid camera but needs to find one homography in each pixel. To ensure a consistent and piecewise

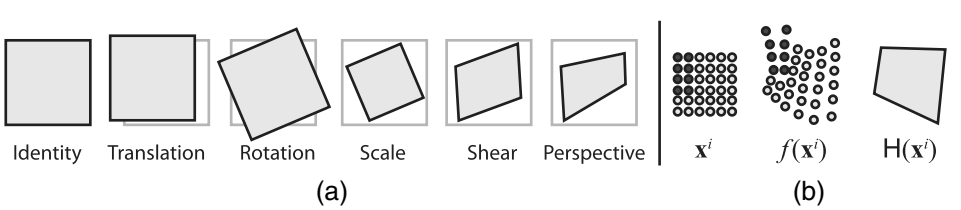

Fig. 3 (a) The effect of identity, translation, rotation, scale, shear, and perspective transformations applied to a quad. Edge-aware moving least-squares estimation of a homography $\mathbf{M}(\mathbf{x})$ from a set of points $\mathbf{x}^i$ undergoing a flow f. (b) Note how pixels from different image content (dark pixels) that undergo a different transformation are not affecting the estimation.

smooth output we combine a regularizing smooth kernel with an edge-aware component w [Eq. (2)].

We implement the entire estimation in parallel over all pixel locations using graphics hardware (GPUs) allowing us to estimate the homography field in less than 3 s for a high-definition image.

3.3 *Transformation Decomposition*

For perceptual scaling the per-pixel transformation $\mathbf{M}$ is decomposed into multiple elementary transformations: translation ($\mathbf{e}_\mathrm{t}$), rotation (e_r), uniform scaling (e_s), aspect ratio change (e_a), shear ($\mathbf{e}_\mathrm{h}$), and perspective ($\mathbf{e}_\mathrm{p}$) (cf. Fig. 4). The relative difficulty of each transformation will later be determined in a perceptual experiment (Sec. 3.5).

We assume that our transformations are the result of a 2-D transformation followed by a perspective transformation. This order is arbitrary, but we decided for it, as it is closer to usual understanding of transformations of 3-D objects in a 2-D world. This is motivated by the fact that it is more natural to imagine objects to live in their (perspective) space and move in their 3-D oriented plane before being projected to the image plane than to understand them as 2-D entities undergoing possibly complex nonlinear and nonrigid transformations in the image plane.

The decomposition happens independently for the matrix $\mathbf{M}$ at every pixel location. As $\mathbf{M}$ is unique up to a scalar, we first divide it by one element, which is chosen to be m_{33}. In the next five steps, each elementary component T will be found first by extracting it from $\mathbf{M}$, and then removing it from $\mathbf{M}$ by multiplying with T^{-1}.

First, perspective is extracted by computing horizontal and vertical focal length as $\mathbf{d}_p = (\mathbf{M}_\mathrm{a}^\mathrm{T})^{-1} \cdot (m_{31}, m_{32}, 0)$ where $\mathbf{M}_\mathrm{a}$ is the affine part of $\mathbf{M}$. The multiplication removes dependency of $\mathbf{d}_p$ on other transformations in $\mathbf{M}$. To define the perceptual measure of the elementary transformation corresponding to the perspective change, we later convert the focal length into the x-axis and y-axis field of view $\mathbf{e}_\mathrm{p} = 2\arctan(\mathbf{d}_p/2)$ expressed in radians. To remove the perspective from $\mathbf{M}$, we multiply it by the inverse of a pure perspective matrix in the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mathbf{d}_{p,y} & \mathbf{d}_{p,y} & 1 \end{pmatrix}.$$

Second, a 2-D vector of translation transformation $\mathbf{e}_\mathrm{t} = (m_{13}, m_{23})$ in visual angle degrees is found. It is removed from $\mathbf{M}$ by multiplying with the inverse of a translation matrix.

Next, we find the rotation transformation corresponding to the angle $e_\mathrm{r} = \arctan 2(m_{21}, m_{11})$ in radians and remove it by multiplying with an inverse rotation matrix.

2-D scaling power is recovered as $\mathbf{d}_s = \log(m_{11}, m_{22})$ and removed from the matrix. For the purpose of later perceptualization, we define a uniform scaling $e_\mathrm{s} = \max(|\mathbf{d}_{s,x}|, |\mathbf{d}_{s,y}|)$ and a change of aspect ratio $e_\mathrm{a} = |\mathbf{d}_{s,x} - \mathbf{d}_{s,y}|$. The assumption is that anisotropic scaling requires more effort to undo than simple isotropic size change and two separate descriptors are, therefore, needed.

The last component is shear defined by a scalar angle as $e_\mathrm{h} = \arctan(m_{12})$ in radians.

3.4 *Transformation Field Entropy*

Definition Ease and difficulty of dealing with transformations does not only depend on the type and magnitude of a transformation on its own but also as it is often the case in human perception it depends on a context. Compensating for one large coherent translation might be easy compared to compensating for many small and incoherent translations. We model this effect using the notion of transformation entropy of an area in an elementary transformation field. Transformation entropy is high if many different transformations are present in a spatial area, and it is low if it is uniform. Note how entropy is not proportional to the magnitude of transformations in a spatial region but to the incoherence in their probability distribution.

We define the transformation entropy H of an elementary transformation at location $\mathbf{x}$ in a neighborhood s using standard entropy equation as

$$H(\mathbf{x}, s) = -\int_\Omega p(\omega|\mathbf{x}, s) \log\, p(\omega|\mathbf{x}, s) \mathrm{d}\omega,$$

where $p(\omega|\mathbf{x}, s)$ is the probability of observing transformation ω in a neighborhood of size s around $\mathbf{x}$. The type of Ω and ω depends on the type of elementary transformation. It is the real plane for translations, scaling, shear, and perspective and the real circle with toroidal wrap-around for rotation. Examples of high and low transformation entropy are shown in Fig. 5.

The probability distribution $p(\omega|\mathbf{x}, s)$ of elementary transformations at neighborhood $\mathbf{x}$, s has to be computed using density estimation, i.e.,

$$p(\omega|\mathbf{x}, s) = \int_{\mathbb{R}^2} K[\omega - t(\mathbf{y})] \mathrm{d}\mathbf{y},$$

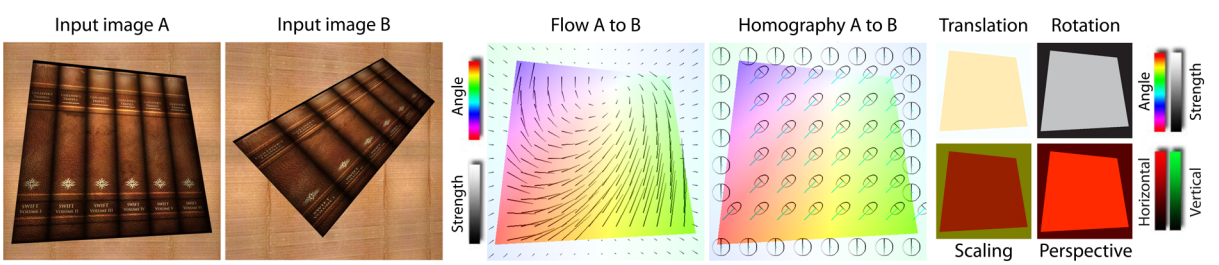

Fig. 4 Conversion of an image pair into elementary transformations. Optical flow from A to B (polar representation) is fitted by local homography matrices (here locally applied on ellipses for illustration) to get five elementary transformations (shear left out for simplicity) having eight channels in total.

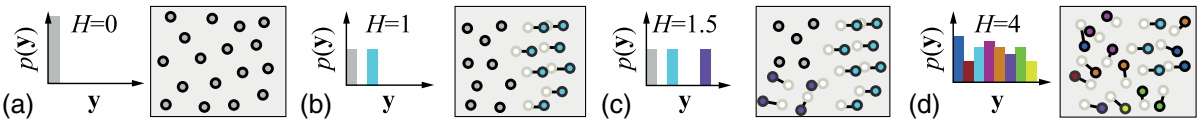

Fig. 5 Entropy for different transformation fields applied to circular items. One transformation maps to one color. (a) For identity, the histogram has a single peak and entropy is 0. (b) For two transformations the histogram has two peaks and entropy is larger. (c) For three transformations, there are three peaks. Entropy is even larger. (d) In increasingly random fields the histogram flattens and entropy increases.

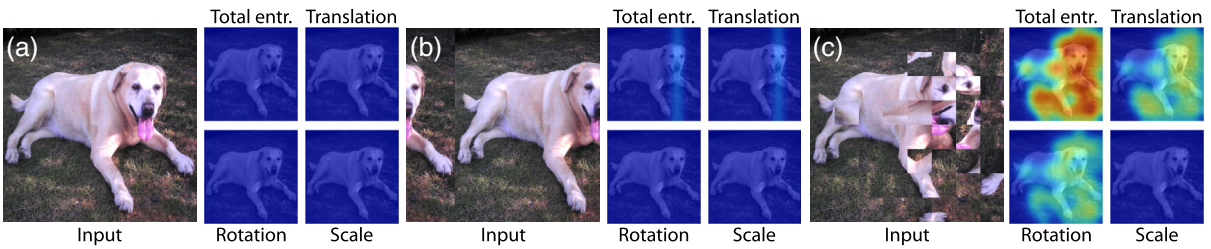

Fig. 6 Entropy of an image under a puzzle-like flow field. (a) No-transformation yields zero entropy. (b) Low entropy is produced by two coherent transformations despite a high average flow magnitude (≈192 px). (c) Transformations with increasing incoherence due to pieces that rotate and swap lead to horizontally increasing entropy despite of a low flow magnitude (≈32 px).

where K is an appropriate kernel such as the Gaussian and $t(\mathbf{x}) \in \Omega$ is a field of elementary transformations of the same type as ω.

Depending on the size of the neighborhood s, entropy is more or less localized. If the neighborhood size is varied, entropy changes as well, resulting in a scale space of entropy,[37] studied in computer vision as an image structure representation. For our purpose, we pick the entropy scale space maximum as the local entropy of each pixel and do not account for the fact at what scale it occurred. The difficulty of transformations was found to sum linearly.[11] For this reason, we sum the entropy of all elementary transformation into a single scalar entropy value.

The decomposition into elementary transformations is the key to the successful computation of entropy; without it, a rotation field would result in a flat histogram as all directions are presented. This would indicate a high entropy, which is wrong. Instead, the HVS would explain the observation using very little information: a single rotation with low entropy.

3.4.1 *Implementation*

In the discrete case, the integral to compute the entropy of the pixel $\mathbf{x}$ becomes

$$\hat{H}(\mathbf{x}) = -\sum_{j=0}^{n_\mathrm{b}} \sum_{k \in \mathcal{N}(s)} K(t_k - \omega_j) \log \sum_{k \in \mathcal{N}(s)} K(t_k - \omega_j), \quad (4)$$

where ω_j is the center of the j'th bin. The inner sums compute the probability of the j'th transformation, essentially by inserting the k'th transformation from a neighborhood $\mathcal{N}(s)$ of size s into one of $n_\mathrm{b} = 32$ bins. An example result is shown in Fig. 6.

Due to the finite size of our n_b histogram bins and the overlap of the Gaussian kernel K, we systematically overestimate the entropy; even when only a single transformation is present, it will cover more than one bin, creating a nonzero entropy. To address this, we estimate the bias in entropy due to a single Dirac pulse and subtract it. We know that 0.99 of the area under a Gaussian distribution is within 3.2 standard deviations σ. That means that a conservative estimate of response to a Dirac pulse is a uniform distribution of the value between 3.2σ bins. That yields the entropy $H_\mathrm{bias} \approx -3.2\sigma(1/3.2\sigma)\log(1/3.2\sigma) = -\log(1/3.2\sigma)$. For our $\sigma = 0.5$ this evaluates to $H_\mathrm{bias} = 0.2$. We approximate the entropy by subtracting this value

$$H(\mathbf{x}) = \hat{H}(\mathbf{x}) - H_\mathrm{bias}. \quad (5)$$

Computing the entropy [Eq. (5)] in a naïve way would require us to iterate a large neighborhoods $\mathcal{N}(s)$ (up to the entire image) for each pixel $\mathbf{x}$ and every scale s. Instead, we use smoothed local histograms[38] for this purpose. In the first pass, the 2-D image is converted into a 3-D image with n layers. Layer i contains the discrete smooth probability of that pixel taking this value. Histograms of larger areas as well as their entropy can now be computed in constant time, by constructing a pyramid on the histograms.

3.5 *Perceptual Scaling*

All elementary transformations as well as the entropy are physical values and need to be mapped to perceptual qualities. Psychological experiments indicate that elementary transformations such as translation, rotation, and scaling require time (or effort) that is close to linear in the relevant x-axis variable[11,13,17,18,39] and that the effect of multiple elementary transformations is additive.[13,18] A linear relation was also suggested for entropy in Hick's law.[30] Therefore, we scale elementary transformation and entropy using a linear mapping (Fig. 7) and treat them as additive.

3.5.1 *Transformation*

To find the scaling, an experiment was performed similar to the one that Shepard and Metzler[11] conducted for rotation but extended to all elementary transformations as defined in

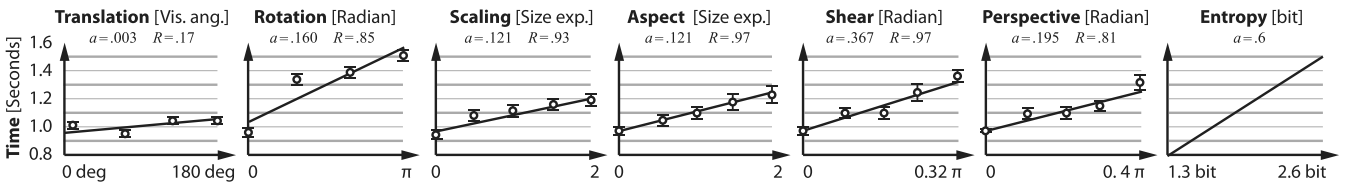

Fig. 7 Perceptual scaling: *x*-axis: increasing elementary transformations and entropy. *y*-axis: increasing difficulty/response time.

Sec. 3.3, including shear and perspective. Objective of the experiment is to establish a relationship of transformation strength and difficulty, measured in response time increase in the mental transformation tasks.

Subjects were shown two abstract 2-D images (Fig. 8) generated from patterns in Fig. 9. The two patterns were either different or identical. One out of the two patterns was transformed using a single elementary transformation of intensity e_i as listed in the upper part of Table 1. Subjects were asked to indicate as quickly as possible if the two patterns are identical by pressing a key. Auditory feedback was provided to indicate if the answer was correct. The time $t(e_i)$ until the response was recorded for all correct answers where the two patterns were identical up to a transformation. The choice of pattern to transform (left or right), the elementary transformation i and its magnitude e_i were randomized in each trial.

Twenty-one subjects (17 M/4 F) completed 414 trials of the experiment in three sessions. For each elementary transformation, we fit a linear function to map strength to response time (Fig. 7). We found a good fit of increasing linear functions of x for all transformations except translation. See Table 1 for the derived model functions. We do not have a definitive answer, why the correlation with translation is lower than for other transformations. A hypothetical explanation is that eye motions can be used to mechanically compensate for translation without mental effort while there is no anatomical option to compensate for rotation, scaling, and so on. An improved design could employ other ways of multiplexing stimuli, e.g., across time, to identify how much translation is cofounded by mental or mechanical aspects, respectively. This agrees with findings for rotation[11] or scaling,[12] and our different bias or slope is likely explained by the influence of stimulus complexity also found by Shepard and Metzler.[11]

3.5.2 *Entropy*

We assume the effect of entropy can be similarly measured as in the task of Hick,[30] where a logarithmic relationship between the number of choices (blinking lamps) and the response time (verbal report of count) was found. He reports a logarithmic time increase with a slope of 0.6 when comparing a visual search task with 10 choices to a single choice-task. The negative logarithm of the inverse number of choices with equal probability is proportional to entropy, so entropy can be directly used for scaling (Fig. 7). We define the response time function of entropy as $t(H) = 0.6H + 0.998$, where the bias constant was gained as a mean response time for zero transformation case in the mental transformation experiment (Table 1), and it is only reported for completeness as only the slope is relevant for our applications.

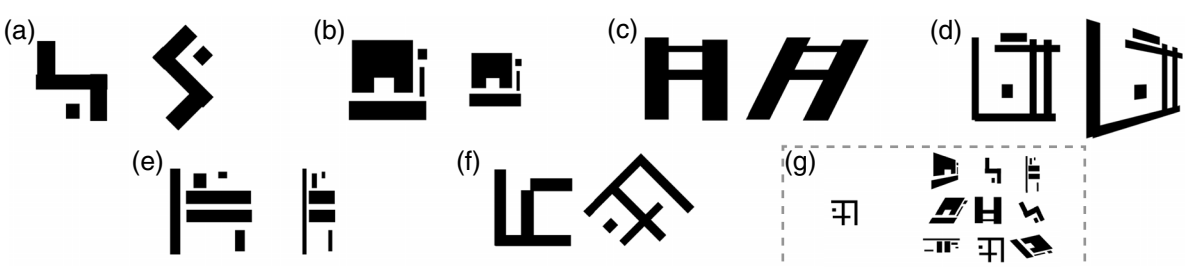

Fig. 8 Some of the stimuli used in our perceptual scaling experiment (Sec. 3.5). The same stimulus under (a) 135 deg rotation requiring 1.4 s mental transformation to undo, (b) 50% scaling (1.1 s), (c) 0.15 π rad shear (1.2 s), (d) 0.15 π rad perspective (1.05 s), and (e) 50% aspect ratio change (1.1 s). (f) A counter example of two different stimuli. (g) A hypothetical extension of our experiment for measuring of the entropy factor as a detection time for an identical stimulus in a growing set of masking objects.

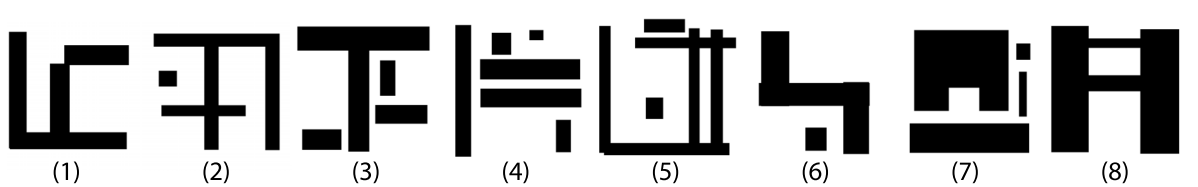

Fig. 9 All eight patterns used to generate stimuli for our perceptual scaling experiment (Sec. 3.5).

Table 1 Results of our perceptual scaling experiment (Sec. 3.5). Input domain corresponds to the range of the transformation parameter presented to users. The response time functions were fitted to our data for elementary transformations and theoretically derived for entropy (Fig. 7). Refer to Secs. 3.3 and 3.4 for the definition of input variables and their units.

Transformation	Input domain	Response time fit	R^2
Translation	$e_t \in [0, 180]$ deg	$t(e_t) = 0.00265e_t + 0.987$	0.171
Rotation	$e_r \in [0, \pi]$ rad	$t(e_r) = 0.00280e_r + 1.053$	0.846
Scaling	$e_s \in [0, 2]$ log .units	$t(e_s) = 0.12100e_s + 0.999$	0.933
Aspect	$e_a \in [0, 2]$ log .units	$t(e_a) = 0.12100e_a + 0.984$	0.972
Shear	$e_h \in [0, 0.32\pi]$ rad	$t(e_h) = 0.00640e_h + 0.973$	0.968
Perspective	$e_p \in [0, 0.4\pi]$ rad	$t(e_p) = 0.00342e_p + 0.989$	0.805
Entropy	$H \in [0, \infty)$ bits	$t(H) = 0.60000\hat{H} + 0.998$	—

4 Applications

The key applications of our model are an image metric (Sec. 4.1), image saliency (Sec. 4.2) and a measure of motion parallax (Sec. 4.3).

4.1 *Image Difference Metric*

The most direct application of our transformation decomposition and its perceptual scaling is building an image difference metric. Our metric does both compare luminance patterns in corresponding regions of the image and also evaluates the strength and the complexity of the spatial relation between them. This way it accounts for the difficulty that the same matching task would cause to the HVS. In combination with a chosen traditional image metric our perceptual transformation measure works as a threshold elevation factor that modifies visibility of image differences (Figs. 1 and 10).

The inputs are two images g_1 and g_2 and their optical flow f as explained in Sec. 3.2. Initially, the second image g_2 is aligned to the first one using the inverse flow f^{-1}. Next, the images can be compared using an arbitrary image metric (we experiment with DSSIM[6]), with the only modification that occluded pixels are skipped from all computations. As a result, a map $\hat{\mathcal{D}}$ is created that contains abstract visual differences [a unitless quality measure in the range from 0 to 1 for $\hat{\mathcal{D}}(g_1, g_2) = DSSIM(g_1, g_2)$ as used in our examples]. This map does not account for the effect of the transformation strength and entropy while we have seen from our experiments that large or incoherent transformations make comparing two images more difficult.

Next, for every pixel we express the increase of difficulty $d_i = t(e_i) - t(0)$ (response time minus optimal response time, i.e., with no transformation or entropy) due to each elementary transformation: translation (d_t), rotation (d_r), scale (d_s), shear (d_h), and perspective (d_p) as well as the entropy (d_H), resulting in a transformation difficulty factor

$$\delta = (1 + d_t + d_r + d_s + d_h + d_p + d_H)^{-1}. \quad (6)$$

The summation is motivated by the finding that response time besides being linear also sums in a linear fashion[12,13] (if scaling adds one second and rotations adds another one, the total time is 2 s).

Finally, we use δ as a factor masking the otherwise potentially perceivable differences in $\hat{\mathcal{D}}$

$$\mathcal{D}(g_1, g_2) = \delta \cdot \hat{\mathcal{D}}(g_1, g_2). \quad (7)$$

As difficulty is in units of time, the resulting unit is visual difference per time. If the original difference map $\hat{\mathcal{D}}$ differed by three units and was subject to a transformation that increased response time by 1 s (e.g., a rotation by about 180 deg), the difference per unit time is $3/(1+1) = 1.5$, whereas a change increasing response time by 3 s (e.g., a shuffling with high entropy) the difference per unit time is $3/(1+3) = 0.75$. In Fig. 10, we show the outcome of correcting the DSSIM index by considering our measure of transformation strength and entropy.

Image transformations that contain local scaling power larger than 0 (zooming) might reveal details in g_2 that were not perceivable or not represented in the first image g_1. Such differences could be reported as indeed they show something in the second image that was not in the first. However, we decided not to consider such differences as a change from nothing into something might not be a relevant change. This can be achieved by blurring the image g_2 with a blur kernel of a bandwidth inversely proportional to the scaling. Occlusions are handled in the same way: No perceived difference is reported for regions only visible in one image.

4.1.1 *Validation*

We validate our approach by measuring a human performance in perceiving differences in an image pair and analyzing its correlation with transformation magnitude and entropy. Subjects were shown image pairs that differed by a flow field as well as a change in content. Two image pairs show 3-D renderings of 16 cubes with different textures [see Figs. 11(a) and 11(c)]. The transformation between the image pairs included a change of 3-D viewpoint and a variety of 3-D motions for each cube. Larger transformations were chosen on the right side of the image [Fig. 11(e)] and similar trend also applies to the entropy introduced by swapping several of the cubes in the grid [Fig. 11(f)].

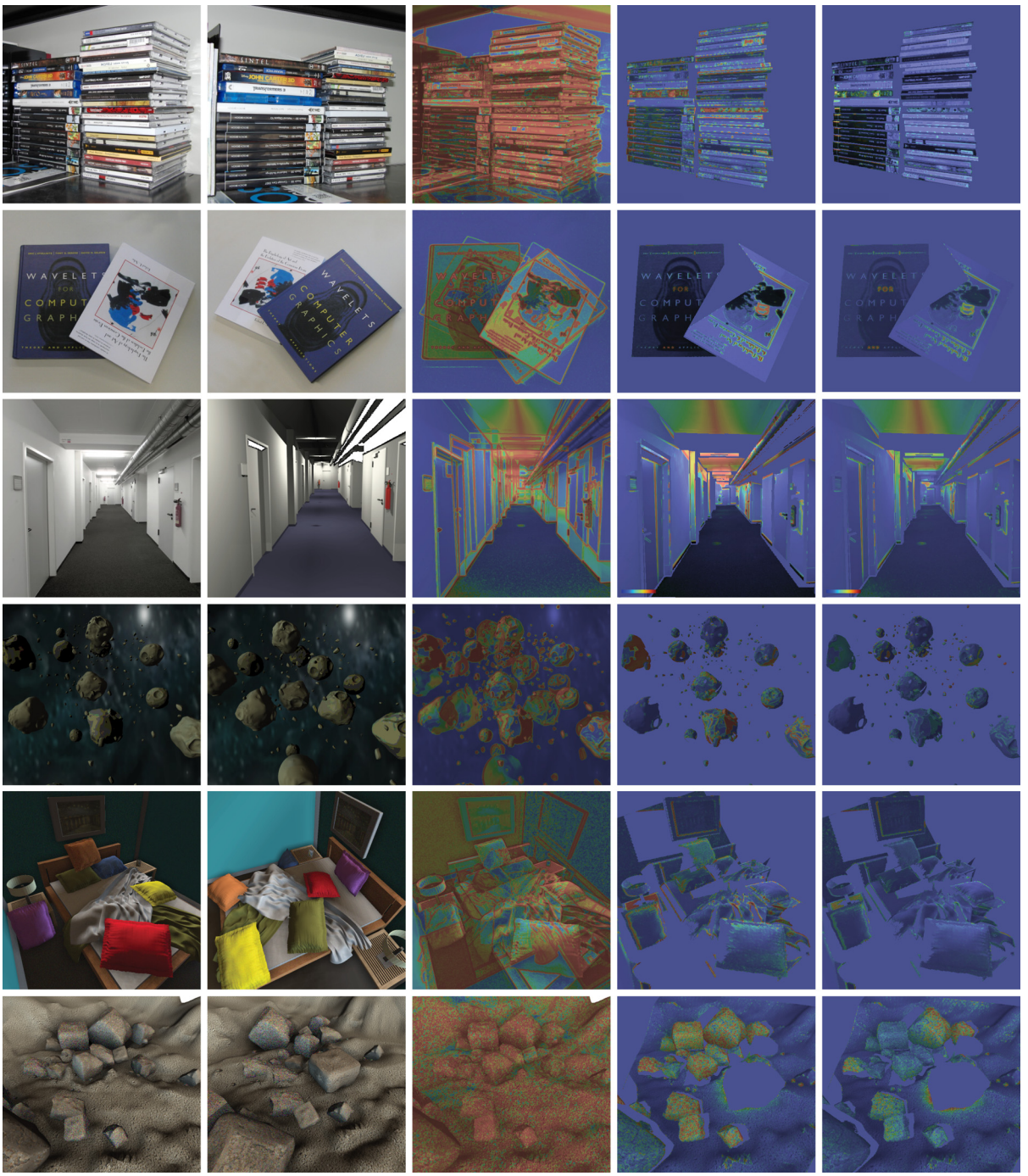

Fig. 10 Results of image difference application using the SSIM index. Rows (top to bottom): (I) A real scene photograph with lot of entropy in the right shuffled CD stack. (II) A simple real scene. (III) A comparison of a photograph and a rerendering of a corridor. (IV) A rendering of flying asteroids. (V) A rendering of a bedroom with some pillows moved. (VI) A rendering of rock landscape with some rocks moved. Columns (left to right): (a, b) Two input images. (c) A naive quality metric without alignment results into false-positive values everywhere. (d) Image alignment itself does not account for high entropy, which would prevent an observer from easily comparing individual objects. (e) Our metric predicts such behavior and marks differences there as less visible.

The images were distorted by adding noise and color quantization to randomly chosen textured cubes, so that the corresponding cubes could differ either by the presence of distortion or their intensity. The intensities of distortions were chosen so that without the geometrical transformation the artifacts are just visible. Ten subjects were asked to mark the cubes that appear different using a 2-D painting interface in an unlimited time. We record the error rate of each object as a relative number of cases the subjects gave wrong answers, i.e., where there was an image difference that they did not mark and where they marked a distortion while there was none [Figs. 11(b) and 11(d)]. Difficult areas have an error rate of 0.5 (chance level) while areas where the subjects were confident have a value as low as 0. The

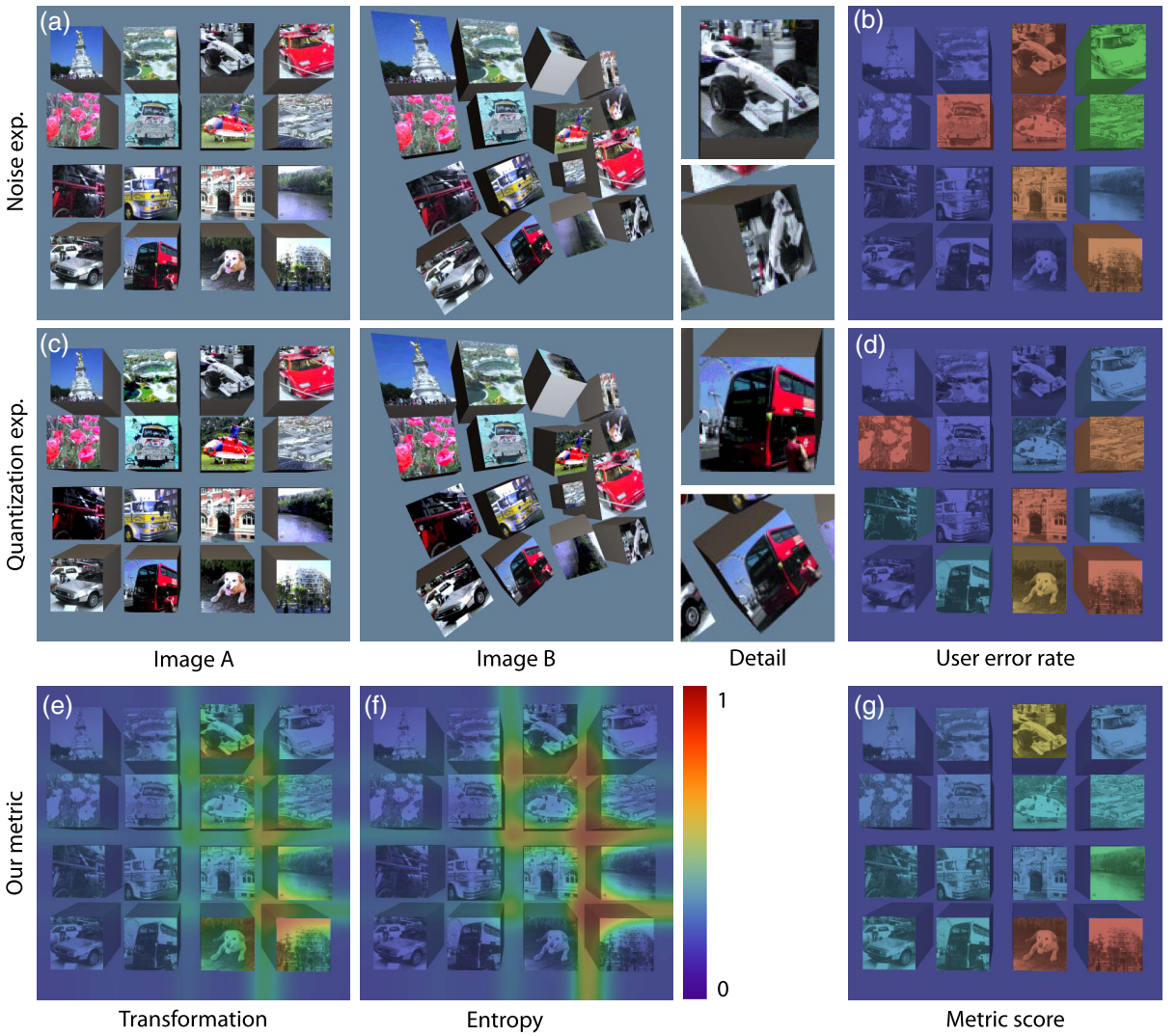

Fig. 11 The validation of our image metric application. An example trial from the user study where a varying degree of geometrical transformation along with a varying distortion intensity is applied to each cube. (a,c) First two rows show the stimuli with the noise and quantization distortion, respectively. (e,f) The third row shows transformation magnitude and entropy measured by our metric for transformation fields shared by both stimuli types. (b,d) The user error rates and (g) our complete difficulty prediction pooled per each cube can be directly compared in the third column. Note that the difficulty metric shown here only measures the transformation influence and therefore is the same for both stimuli. See Sec. 4.1 for the full image metric.

error rate is averaged over all subjects for one distortion and one scene.

Our transformation difficulty metric consists of the transformation magnitude measures and entropy, and we pool it for each cube by averaging to obtain 16 scalar values [Fig. 11(g)]. As the geometrical layout is the same for both types of distortions and each was setup to have a similar visibility, the assumption is that the subjects would find the same geometrical transformations difficult in both cases. That means that error rates in Figs. 11(b) and 11(d) should be similar and we correlate our metric with both of them together.

We analyzed the correlation of the error rate [Figs. 10(b) and 10(d)] and transformation magnitude and entropy [Figs. 10(e)–10(g)] and found an average Pearson's r correlation of 0.701 (Table 2), which is significant according to the t-test with $p < 0.05$. The transformation magnitude has a lower correlation of $r = 0.441$ compared to transformation entropy $r = 0.749$ (significant for $p < 0.05$). This difference could possibly be explained by a design of our experiment. Given an unlimited time, the subjects were eventually able to undo all transformations and resolve all shuffling between the cubes. It may be that the short-term memory requirement made the shuffling problem harder than the one resulting from the transformations. This would increase the importance of entropy for the performance prediction and lead to a higher correlation. A time constrained version of the experiment could answer this question. Another possible explanation points to a relatively low magnitude of transformations applied to our stimuli compared to the entropy introduced by shuffling of many similar cubes. An experiment design with different combinations of both factors could be used to verify this theory. Despite this asymmetry, we conclude that both transformation magnitude and entropy

Table 2 Correlations of image differences from variants of our metric [Figs. 10(e)–10(g)] and average error rate of our study participants [Figs. 10(b) and 10(d) together] as described in Sec. 4.1. Stars* denote significance according to the *t*-test with $p < 0.05$.

Metric	Correlation
Only transformation magnitude	0.441
Only transformation entropy	0.749*
Our full metric	0.701*

correlates with the ability to detect distortions; in the presence of strong or complex transformation, the increase in human detection error can be fit using a linear model.

The final performance of our approach is limited by the image metric used. The correlation of image metrics and quantitative user responses is low and difficult to measure[40] even without transformations. Therefore, the evaluation of the full metric, in particular for suprathreshold conditions, is relegated to future work.

4.1.2 *Discussion*

Here we discuss the relation of our and existing image and video metrics, in particular how they deal with transformations. For a more general survey of image quality metrics we refer to Ref. 6.

Standard image difference (fidelity) metrics, such as per-pixel MSE, peak signal-to-noise ratio, per-patch structure similarity (SSIM) index,[6] or the perception-based visible differences predictor[41] are extremely sensitive for any geometric distortions (Ref. 6, Figs. 1.1 and 3.8). The requirement of perfect image registration is lifted for the pixel correspondence metric,[42] closest distance metric,[43] or point-to-closest-point MSE, which account for edge distances. Natural images can be handled by the complex wavelet CW-SSIM index but mostly small translations can be supported (Ref. 6, Ch. 3.2.3).

Liu and Chen[44] describe a method for JPEG artifact assessment that compares the power spectrum in the frequency domain. Although not aiming for a complete transformation invariance as in our case, some degree of tolerance for an imperfect alignment can be expected. A similar analysis was also demonstrated using wavelets.[45] Zhou et al.[46] combined a comparison of mutual differences between the reference and test image with an analysis of self-similarity within each of the images. Such internal similarity can be better preserved than mutual similarity especially when the complexity of transformation is relatively low. A machine learning approach can be used to improve the adaptability of a metric to various content. Jin et al.[47] train a metric selecting a specialized approach for a content structure and a distortion category. A potential extension would teach the metric to recognize local transformations and enable it to undo them. An application where the spatial alignment of images cannot be taken for granted is a quality assessment for stereoscopic 3-D. Changes in disparity will cause the left and right image to shift, often in spatially nonuniform way. Li et al.[48] showed how the disparity information can be used to undo the stereoscopic projection and merge left and right eye into a cyclopean image where luminance features can be compared in a similar way as in SSIM.

All these approaches model local deformation invariance, which is a low-level (C1 cortical area) process. Our transformation-aware quality metric attempts to compensate for transformations of much larger magnitude, which occurs at higher levels[10] including perspective transformation.

Video quality assessment typically considers the temporal domain as a straightforward extension of 2-D spatial filter banks into 3-D.[49] This precludes any reasonable motion analysis based on its direction and velocity, which requires the optical flow computation. A notable exception is the work of Seshadrinathan and Bovik[50] where the optical flow is derived using 3-D Gabor filters that span both the spatial and the temporal domain in order to evaluate the spatial quality of video frames, as well as the motion quality. Dominant motion increases the perception uncertainty and suppresses distortion visibility, while relative motion can make video degradations easier to notice.[49] Our homography decomposition enables to analyze dominant and relative motion, and our transformation entropy accounts for their local complexity, which we utilize in our transformation-aware quality metric.

The visual image equivalence[1] measures whether a scene's appearance is the same rather than predicting if a difference is perceivable. Perceivably different scenes can result in the same impression, as the HVS compensates for irrelevant changes. Our method can be considered another form of visual equivalence, as we model compensation for transformations. Comparing two aggregates of objects[51] is also related to entropy but goes beyond, if the aggregates differ by more than a transformation, i.e., deletion or insertion of objects.

4.1.3 *Limitations*

Our metric heavily relies on the quality of the optical flow estimation. We analyze the transformation and entropy in the image by fitting homography locally to a small neighborhood. This can potentially lead to an amplification of the noise in the original optical flow, and consequently, overestimation of the transformation and its entropy. A special care has to be given to textureless or occluded regions where the flow estimate is unreliable. Such regions should conservatively be ignored in computing the final metric by setting $\delta = 1$.

Another limitation of our metric is its focus on transformation properties alone. It is not clear how luminance properties, such as local contrast or texture distinctiveness, influence the ability of the HVS to understand the transformation, which has a direct influence on the perceivability of image differences. Our metric could therefore overestimate the performance of the HVS in regions where luminance patterns are confusing for understanding of the motion. The dazzle camouflage is one such example of a luminance pattern, which makes estimation of an object shape and position very difficult.[52]

4.2 *Saliency*

Saliency estimation has a lot of applications in computer graphics (Sec. 2.3). It can substitute for a direct eye tracking and enable image processing optimized for important regions of the image that are more likely to be observed than others.

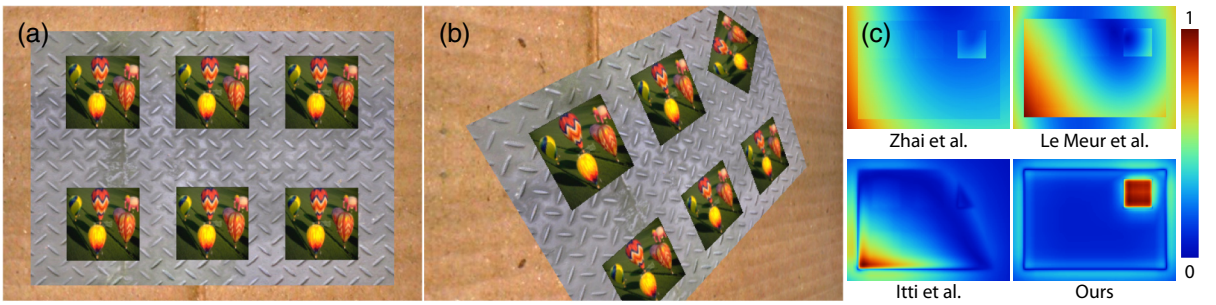

Fig. 12 Transformation-aware saliency for (a) an input image being deformed into image (b). One patch is salient, as it deforms differently. Only transformation saliency is shown. This "differently," however, is nontrivial and can only be detected with a transformation representation that captures the human ability to compensate for certain transformations including perspective, such as ours. Other motion saliency methods do not capture this effect (c), but instead consider other parts more salient, following local variations in the flow.

While some other saliency models also consider motion, our method is unique in decomposing motion into elementary transformations (Fig. 12). This allows for an easier detection of a distinct motion patterns that can be obstructed in the original optical flow. Such unique features easily causes a "pop-out" effect, which attracts a user attention, hence it increases saliency.

Different from common image saliency, our approach takes two instead of one image as an input. It outputs saliency, e.g., how much attention an image region receives. We largely follow the basic, but popular, model of Itti et al.,[25] and replace its motion detection component by our component that detects salient transformations.

First, we compute a feature map for every elementary transformation signal e_i (Sec. 3.3) as

$$E_i(c,s) = |e_i(c) \ominus e_i(s)|, \tag{8}$$

where $\ominus$ is an operator computing contrast between two levels $c \in \{2,3,4\}$ and $s = c + \{3,4\}$ of a Gaussian pyramid for an elementary transformation signal e_i. Typically, $\ominus$ is a simple difference but special care has to be taken for periodicity of angular values in case of rotation. Transformations with vector values, such as translation $\mathbf{e}_t$, are treated as separate entities for each dimension. Six resulting feature maps per each transformation are then combined into corresponding conspicuity maps[25] $\bar{E}_i$ by summation

$$\bar{E}_i = \sum_{c,s} E_i(c,s). \tag{9}$$

Finally, all conspicuity maps are normalized and averaged to get the final motion saliency score

$$\mathcal{S} = \sum_i \mathcal{N}(\bar{E}_i), \tag{10}$$

where $\mathcal{N}(\cdot)$ is the normalization operator by Itty et al.[25]

4.2.1 *Discussion*

We compare our approach for scenes that contain complex elementary transformations to approaches by Le Meur et al.,[53] Zhai and Shah,[54] and Itti et al.[25] (with the original motion detection component) in Fig. 12.

A vast majority of saliency models that handle temporal domain are focused on motion detection with respect to the static environment,[7] but motion pop-out may also arise from nonconsistent relative motion. Therefore the global motion (e.g., due to camera motion) or consistent and predictable object motion should be factored out, and the remaining relative motion is likely to be a strong attention attractor. Along this line, Le Meur et al.[53] derive the global motion in term of an affine transformation using robust statistics and remove it from the optical flow. The remaining outlier flow is compared to its median magnitude as a measure of saliency. Such per-pixel statistics make it difficult to detect visually consistent object transformations, such as rotations, where the variability of the motion magnitude and direction might be high. Zhai and Shah[54] derived local homographies that model different motion segments in the frame. In this work, we compute transformation contrast similar to translation-based motion contrast in Ref. 54, but we perform it for all elementary transformations, and we account for neighboring homographies in a multiresolution fashion, instead considering all homographies at once. This gives us a better locality of transformation contrast. Also, through decomposition into elementary transformations we are able to account for the HVS ability to compensate for numerous comparable (nonsalient) transformation components akin to camera or large object motion and detect highly salient unique motion components. This way, instead of detecting local variations of optical flow, we are able to see more global relations between moving objects (as relative rotation in Fig. 12). The edge-stopping component of homography estimation enables us to find per-pixel boundary of regions with inconsistent motion, which further improves the accuracy of saliency maps. Finally, our saliency model is computationally efficient and can be performed at near-interactive rates.

4.2.2 *Limitations*

Similarly to our image metric application our saliency predictor depends on the availability of a good optical flow estimate. Unreliable and noisy optical flow would reduce the efficiency of our method as an increase of entropy in the image generally lowers the prominence of the salient feature. Note that this makes it difficult to validate our method using standard saliency datasets with ground truth attention data gathered using eye tracking. Such datasets usually do not

contain pairs of images as required by our method. A video input could potentially be used but the reliable optical flow data are missing in existing datasets. We conclude that a saliency validation dataset containing spatial transformations and registration information in the form of optical flow is highly desirable.

4.3 *Motion Parallax*

Motion parallax is defined by relative retinal motion due to a rigid geometrical transformation of the scene during motion of the object or the observer. Analyzing variance of optical flow directly can easily lead to an overestimation of the relative motion since nonlinear transformations, such as rotation, will yield incoherent optical flow. Our transformation decomposition removes this problem by attributing each elementary transformation to its proper magnitude map e_i. This way relative motion can be analyzed more robustly by investigating each map separately.

We propose a measure of motion parallax (Sec. 2.4), which relies on a combination of spatial change of flow (motion contrast) and spatial change of luminance (luminance contrast) (Fig. 13, Middle). First, an elementary transformation is computed for each level of the pyramid. The resulting pyramids contain at every pixel the difference in motion between a pixel and its spatial context of a size that depends on the level. Then, we compute the absolute values of such differences.

First, similar to a Laplacian image pyramid,[55] we build a contrast pyramid $\mathcal{C}_i$ for each elementary transformation e_i (Sec. 3.3)

$$\mathcal{C}_i(j) = |e_i(j) \ominus e_i(j+1)|, \tag{11}$$

where $\ominus$ is the same operator as in Eq. (8) and $e_i(j)$ is j'th level of the Gaussian pyramid of elementary transformation i. Finally, each pyramid is collapsed into a single image, and we sum the resulting images for all elementary transformations

$$\mathcal{Q} = \sum_i \sum_j \mathcal{C}_i(j). \tag{12}$$

High values of $\mathcal{Q}$ indicate a strong parallax effect and low values indicate its absence. We show applications of this measure of parallax in two rendering applications.

4.3.1 *Adaptive parallax occlusion mapping*

Real world objects often exhibit complex surface structures that would be too costly to directly model for computer graphic rendering applications. The huge geometry complexities would quickly surpass the memory capacity or the rendering throughput of any hardware. That is why the objects get simplified and surface details replaced by a single plane. This is efficient but it reduces the realism as both shading and structural details cannot be correctly reproduced [Fig. 14(a)]. The normal mapping[56] uses additional surface raster to encode normal details, which are then used to reconstruct high frequencies of the shading at cost of a small memory and performance overhead [Fig. 14(b)]. The results are convincing for a static scene but adding a motion reveals that the motion parallax cue is completely missing. The parallax occlusion mapping[57] (POM) tackles this issue by adding a surface-displacement map and performing a simple ray tracing to determine visibility and occlusions inside the surface plane [Fig. 14(c)]. Although significantly more costly, this effect is popular in current games as it greatly improves the realism. The importance of this effect has also been noticed in head mounted displays for virtual reality applications where the mismatch between head motion and the lack of motion parallax is strongly objectionable. As the

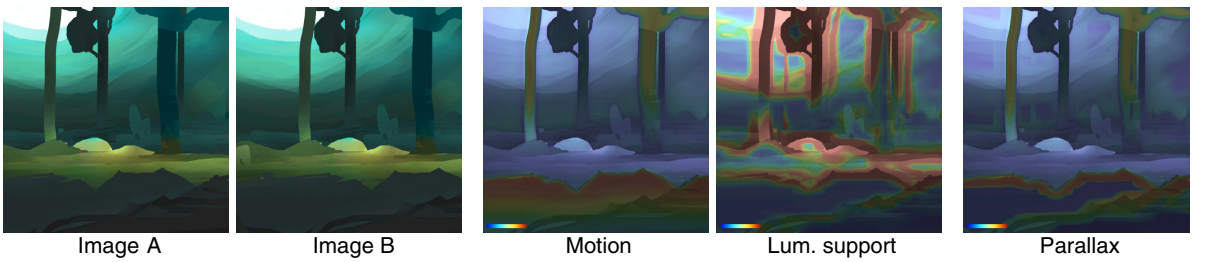

Fig. 13 Motion parallax between images A and B.

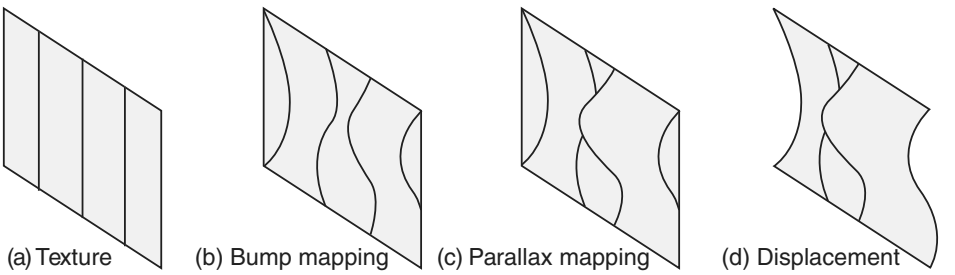

Fig. 14 Examples of surface mapping techniques used in computer graphics. (a) Classic shading cannot reproduce high frequency details that are not part of the geometry. (b) Normal mapping introduces detail shading variation. (c) POM allows for motion parallax inside the object. (d) Displacement mapping additionally supports changes in the object outline.

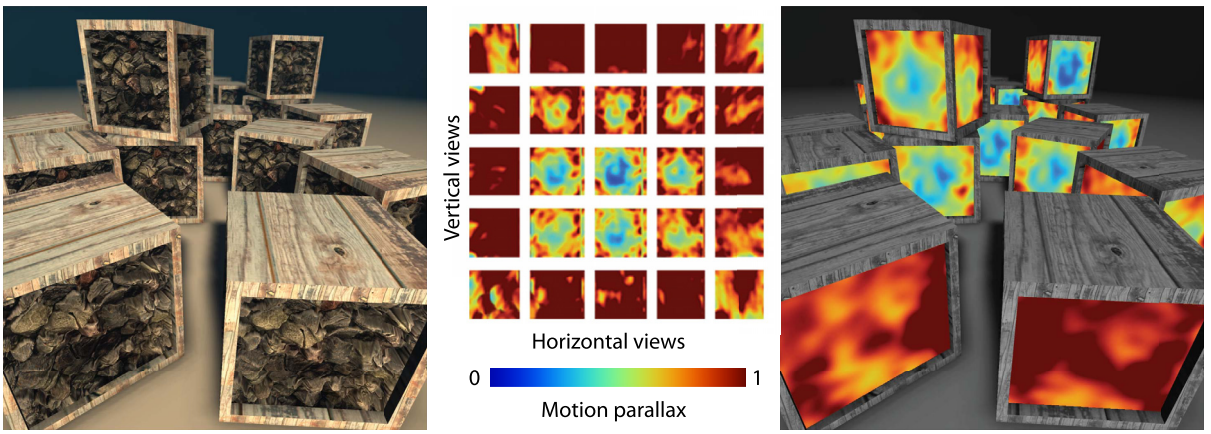

Fig. 15 Two frames of an image sequence where we selectively enable POM based on a preprocessing of the displacement map that predicts for which view and illumination a parallax effects is perceivable, decreasing the number of POM ray casting steps by up to 50% and rendering time from 13.5 to 8.9 ms.

computation happens inside the originally flat surface, its outline cannot be modified. That means that the shape of the object can still reveal the underlying simplified model. A remedy for this is provided by the displacement mapping, which uses HW capability of modern GPUs to generate a detail geometry matching the height map on the fly, therefore, without memory footprint [Fig. 14(d)]. As both of the later techniques are quite costly to compute, their use should be driven by a benefit that they bring to the user. We demonstrate how such a decision can be supported by our method on the case of POM.

In case of the POM, the height field needs to be ray-traced for every pixel, limiting its use in interactive applications, especially on low-power, i.e., mobile devices. Using our approach we can detect where a certain displacement map will actually result in a perceivable motion parallax for a certain view direction when this view is slightly changed and adjust the quality of the effect per-pixel, saving considerable compute time and bandwidth (Fig. 15). To this end, we prerender the displacement map, including texturing for all view directions and compute the motion parallax for a differential motion along each spherical direction for all view directions (refer to the inset in Fig. 15). The result is a lookup function that indicates how much a pixel benefits from POM from a certain view or not. At runtime, we look up the value for a given view direction, and adjust the size of a ray-tracing step for each pixel. That modifies the number of iterations required for evaluation of the effect and, therefore, the required computational time.

4.3.2 *Motion parallax-based viewpoint selection*

Motion parallax is an effective depth cue,[26] and cinematographers know how to use it to convey the layout of a scene. To our knowledge however, there is not yet an automated way to pick the right camera or object motion, such that the resulting motion parallax is most effective. Consequently, casual users that need to place a camera will have difficulties selecting it effectively.

Using our approach, we can derive an extended view point+motion selection approach (Fig. 16) that, given a 3-D

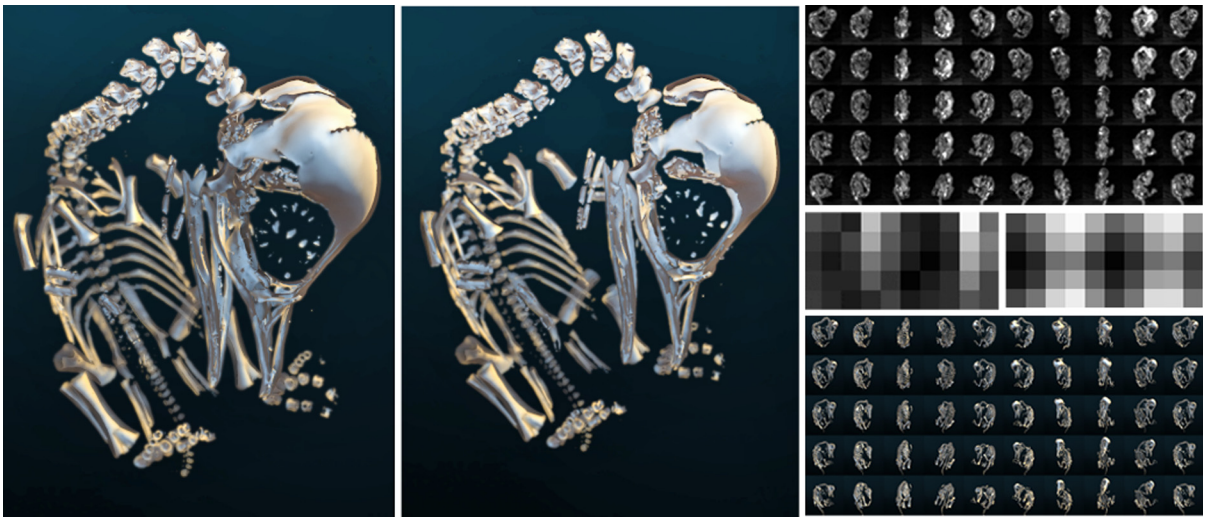

Fig. 16 Selection of two views for an anatomical dataset (UTCT, U Texas), where motion parallax when flipping the two images at the right speed reveals most about the depth layout. The right column shows the spatially varying motion parallax (top), the integrated motion parallax and common viewpoint preference (center), and the images for all view directions (bottom).

scene, picks a view direction and a change of view position, such that the resulting image pair features optimal motion parallax. The image pair can directly be used in a stereo flip animation or as two key frames of a very slow camera motion, akin to the Ken Burns' effect in 2-D (Ref. 58, page 512). Optionally, the view position can be fixed, and only the direction of change is suggested, or vice versa. To compute the pair, we use the same approach as for POM. We densely sample the motion parallax of the entire image for the 2-D set of all view directions and their ϕ and θ derivatives in spherical coordinates. We return the pair where the motion parallax is maximal, optionally combined with other viewpoint criteria.

5 Conclusion and Future Works

We propose a model of human perception of transformations between a pair of images. The model converts the underlying optical flow into a field of homographies, which is further decomposed into elementary transformations that can be perceptually scaled and allows the notion of transformation entropy. Our model enables for the first time a number of applications. We extended perceptual image metrics to handle images that differ by a transformation. We extend visual attention models to detect conspicuous relative object motion, while ignoring predictable motion such as due to view changes or consistent object motion. Finally, we provide a measure of motion parallax based on optical flow and demonstrate the utility of this measure in rendering applications to steer adaptive POM and viewpoint selection.

Our transformation-aware perceptual scaling may have other interesting applications, which we relegate as future work. In image change blindness,[59] the same view has been considered so far, and our approach could be beneficial to predict the increased level of difficulty in the visual search task due to perspective changes. Also, the concept of visual equivalence[1] can be extended to handle different scene views, as well as minor deformations of the object geometry and their relocation. Our quality metric could be applicable to rephotography and rerendering[3] allowing for a better judgement of structural image differences while ignoring minor misregistration problems. This is also the case in image retargeting, where all image distance metrics such as SIFT flow, bidirectional similarity, or earth mover's distance[4] account for some form of the energy cost needed to transform one image into another. While semantic and cognitive elements of image perception seem to be the key missing factors in those metrics, it would be interesting to see whether our decomposition of the deformation into elementary transformations and perceptual scaling of their magnitudes could improve the existing energy-based formulations.

In the end of the day the basic question is: "what is an image?" In most cases, "image" does not refer to a matrix of physical values but refers the mental representation of a scene. This mental representation is created by compensating for many variations in physical appearance. The ability to compensate for transformation as well as its limitations are an important part of this process and has been modeled computationally in this work.

References

1. G. Ramanarayanan et al., "Visual equivalence: towards a new standard for image fidelity," *ACM Trans. Graph.* **26**(3), 76 (2007).
2. G. Meyer et al., "An experimental evaluation of computer graphics imagery," *ACM Trans. Graph.* **5**(1), 30–50 (1986).
3. S. Bae, A. Agarwala, and F. Durand, "Computational rephotography," *ACM Trans. Graph.* **29**(3), 24 (2010).
4. M. Rubinstein et al., "A comparative study of image retargeting," *ACM Trans. Graph.* **29**(6), 160 (2010).
5. J. J. Gibson, *The Perception of the Visual World*, Houghton Mifflin (1950).
6. Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers (2006).
7. A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 185–207 (2013).
8. I. P. Howard and B. J. Rogers, *Perceiving in Depth*, I. Porteous, Toronto (2012).
9. J. Koenderink, "Optical flow," *Vis. Res.* **26**(1), 161–179 (1986).
10. G. A. Orban et al., "First-order analysis of optical flow in monkey brain," *Proc. Natl. Acad. Sci. U. S. A.* **89**(7), 2595–2599 (1992).
11. R. Shepard and J. Metzler, "Mental rotation of three dimensional objects," *Science* **171**(3972), 701–703 (1971).
12. C. Bundesen, A. Larsen, and J. Farrel, "Mental transformations of size and orientation," *Attention and Performance* **IX**, 279–294 (1981).
13. R. Sekuler and D. Nash, "Speed of size scaling in human vision," *Science* **27**(2), 93–94 (1972).
14. L. Cooper, "Demonstration of a mental analog of an external rotation," *Percept. Psychophys.* **19**(4), 296–302 (1976).
15. J. E. Cutting, "Rigidity in cinema seen from the front row, side aisle," *J. Exp. Psych.: Human Percept. Perform.* **13**(3), 323 (1987).
16. D. Vishwanath, A. R. Girshick, and M. S. Banks, "Why pictures look right when viewed from the wrong place," *Nat. Neurosci.* **8**(10), 1401–1410 (2005).
17. C. Robins and R. Shepard, "Spatio-temporal probing of apparent motion movement," *Percept. Psychophys.* **22**(1), 12–18 (1977).
18. C. Bundesen and A. Larsen, "Visual apparent movement: transformations of size and orientation," *Perception* **12**, 549–558 (1983).
19. J. T. Todd, "The visual perception of 3D structure from motion," in *Perception of Space and Motion*, W. Epstein and S. Rogers, Eds., pp. 201–226, Academic Press, Cambridge, Massachusetts (1995).
20. R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer, London (2011).
21. R. A. Eagle, M. A. Hogervorst, and A. Blake, "Does the visual system exploit projective geometry to help solve the motion correspondence problem?," *Vis. Res.* **39**(2), 373–385 (1999).
22. F. Liu et al., "Content-preserving warps for 3D video stabilization," *ACM Trans. Graph.* **28**(3), 44 (2009).
23. C. Liu, J. Yuen, and A. Torralba, "SIFT flow: dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(5), 978–994 (2011).
24. T. Igarashi, T. Moscovich, and J. F. Hughes, "As-rigid-as-possible shape manipulation," *ACM Trans. Graph.* **24**(3), 1134–41 (2005).
25. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998).
26. B. Rogers and M. Graham, "Motion parallax as an independent cue for depth perception," *Perception* **8**(2), 125–134 (1979).
27. J. Y. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.* **3**(5), 625–638 (1994).
28. R. M. Bevenesee, *Maximum Entropy Solutions to Scientific Problems*, Prentice-Hall (1993).
29. E. Grossman, "Entropy and choice time: the effect of frequency unbalance on choice-response," *Q. J. Exp. Psychol.* **5**(2), 41–51 (1953).
30. W. E. Hick, "On the rate of gain of information," *Q. J. Exp. Psychol.* **4**(1), 11–26 (1952).
31. J. P. W. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Trans. Med. Imaging* **22**(8), 986–1004 (2003).
32. P.-P. Vázquez et al., "Automatic view selection using viewpoint entropy and its applications to image-based modelling," *Comput. Graphics Forum* **22**(4), 689–700 (2003).
33. S. Gumhold, "Maximum entropy light source placement," in *IEEE Visualization*, pp. 275–282 (2002).
34. T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 41–48 (2009).
35. R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(6), 580–593 (1997).
36. C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth Int. Conf. on Computer Vision, 1998*, pp. 839–846 (1998).
37. M. Ferraro, G. Boccignone, and T. Caelli, "On the representation of image structures via scale space entropy conditions," *IEEE Trans. Pattern Anal. Mach. Intell.* **21**(11), 1199–1203 (1999).
38. M. Kass and J. Solomon, "Smoothed local histogram filters," *ACM Trans. Graph.* **29**(4), 100 (2010).
39. J. Huttenlocher and C. C. Presson, "Mental rotation and the perspective problem," *Cog. Psych.* **4**(2), 277–299 (1973).

40. M. Čadk et al., "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Trans. Graph.* **31**(6), 147:1–147:10 (2012).
41. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," *Proc. SPIE* **1666**, 2 (1992).
42. M. S. Prieto and A. R. Allen, "A similarity metric for edge images," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(10), 1265–1273 (2003).
43. K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge detector evaluation using empirical ROC curves," *Comput. Vision Image Understanding* **84**(1), 77–103 (2001).
44. B. Liu and H. Chen, "Quality assessment for jpeg images based on difference of power spectrum distribution," *Front. Optoelectron.* **8**(4), 419–423 (2015).
45. S. Rezazadeh and S. Coulombe, "A novel discrete wavelet transform framework for full reference image quality assessment," *Signal Image Video Process.* **7**(3), 559–573 (2013).
46. F. Zhou et al., "Image quality assessment based on inter-patch and intra-patch similarity," *PLoS One* **10**(3), e0116312 (2015).
47. L. Jin, K. Egiazarian, and C.-C. J. Kuo, "Perceptual image quality assessment using block-based multi-metric fusion (BMMF)," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '12)*, pp. 1145–1148, IEEE (2012).
48. F. Li et al., "Full-reference quality assessment of stereoscopic images using disparity-gradient-phase similarity," in *2015 IEEE China Summit and Int. Conf. on Signal and Information Processing (ChinaSIP '15)*, pp. 658–662, IEEE (2015).
49. Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am. A* **24**(12), B61–B69 (2007).
50. K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.* **19**(2), 335–350 (2010).
51. G. Ramanarayanan, K. Bala, and J. A. Ferwerda, "Perception of complex aggregates," *ACM Trans. Graph.* **27**(3), 60 (2008).
52. N. E. Scott-Samuel et al., "Dazzle camouflage affects speed perception," *PLoS One* **6**, e20233 (2011).
53. O. Le Meur et al., "A spatio-temporal model of the selective human visual attention," in *IEEE Int. Conf. on Image Processing (ICIP '05)*, pp. 1188–1191 (2005).
54. Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. of the 14th ACM Int. Conf. on Multimedia (MM '06)*, pp. 815–824 (2006).
55. P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.* **31**(4), 532–540 (1983).
56. J. Cohen, M. Olano, and D. Manocha, "Appearance-preserving simplification," in *Proc. of the 25th Annual Conf. on Computer Graphics and Interactive Techniques*, pp. 115–122, ACM (1998).
57. N. Tatarchuk, "Dynamic parallax occlusion mapping with approximate soft shadows," in *Proc. of the 2006 Symp. on Interactive 3D Graphics and Games (I3D '06)*, pp. 63–69 (2006).
58. T. Green and T. Dias, *Foundation Flash CS5 For Designers*, IT Pro, Apress (2010).
59. L. Q. Ma et al., "Change blindness images," *IEEE Trans. Vis. Comp. Graph.* **19**(11), 1808–1819 (2013).

Petr Kellnhofer is a PhD candidate under a joint supervision of Prof. Karol Myszkowski and Prof. Hans-Peter Seidel at MPI Informatik, Saarbücken, Germany, since 2012. His research interests cover application of perception to computer graphics with a focus on stereoscopic 3-D. During his PhD he visited the group of Prof. Wojciech Matusik at MIT CSAIL, where he investigated eye-tracking methods and their applications.

Tobias Ritschel is a senior lecturer (associate professor) in the VECG group at the University College London. His interests include interactive and nonphotorealistic rendering, human perception, and data-driven graphics. He received the Eurographics PhD dissertation award in 2011 and Eurographics Young Researcher Award in 2014.

Karol Myszkowski is a tenured senior researcher at the MPI Informatik, Saarbücken, Germany. From 1993 till 2000, he served as an associate professor in the Department of Computer Software at the University of Aizu, Japan. His research interests include global illumination and rendering, perception issues in graphics, high dynamic range imaging, and stereo 3-D.

Hans-Peter Seidel is a scientific director and chair of the Computer Graphics Group at the MPI Informatik and a professor of computer science at Saarland University. In 2003, he received the Leibniz Preis, the most prestigious German research award, from the German Research Foundation (DFG). He is the first computer graphics researcher to receive such an award.